RECSM

Research and Expertise Centre for Survey Methodology

# Evaluation of the quality and invariance of survey questions

**Willem Saris**

RECSM Working Paper Number 39

June 2014

http://www.upf.edu/survey/_pdf/RECSM_wp039.pdf

**Abstract**

At several occasions we have been asked why our research with respect to quality and invariance of survey questions has been based on analyses of product-moment (Pearson) correlations and not on polychoric and polyserial correlations or latent trait models for categorical variables. In this research note a description will be given of the possible options , their advantages and disadvantages based on Monte Carlo experiments, real experiments and a meta-analysis. After that we will indicate the empirical arguments that brought us to the choice we have made.

**Basic reasons for evaluating survey questions**

1. A distinction should be made between concepts by postulation (CP) and concepts by intuition (CI). The latter can be measured by a single question. The former are based on several concepts by intuition (Northrop 1947).
2. We concentrate on the distribution and relationships between the CI. The consequences for the CP derive from the way we treat the CI
3. There is a difference between the CI and the observed responses for different reasons: categorization (grouping and transformation errors), random errors and systematic errors.
4. These errors can considerably change the estimates of relationships between the observed variables (Duncan and Goldberger 1971). The errors can also be different across countries which makes comparison of relationships between variables across countries impossible without correction for measurement errors. Therefore the relationships between the concepts cannot be studied without correction for measurement errors (Saris and Gallhofer 2007).
5. If one corrects for measurement errors then the further analysis between the different CI and between CI and CP can be done free of errors. In that case comparison of the relationships across countries is possible as well (Saris and Gallhofer 2007))
6. If the correlations between the CI are corrected for known measurement errors or the quality of the questions then one can also study in how far the used CI are invariant with respect the relationship with the CP of interest across countries, i.e. that the understanding of the questions is functional equivalent.
7. If one knows the qualities of the questions that are seen as indicators for CP then one can also estimate the quality of the composite score for CP
8. So the fundamental questions are: what are the quality indicators for survey questions and how can these quality indicators be estimated?

**Definition of quality indicators**

1. Lord and Novick (1968) defined reliability coefficient as the correlation between the true score and the observed score, i.e. the reliability ($r^2$) = 1- the error variance
2. It has been argued convincingly by Campbell and Fiske (1959) and others that one can expect in any measure random errors and systematic errors. There are many possible sources for systematic errors. Campbell and Fiske directed the attention to the fact that the use of the same method can lead to different reactions of people which are stable across questions and therefore create correlations that have nothing to do with

the relationships between the CI. This phenomenon is called "common method variance".

3. Saris and Andrews (1991) defined the validity coefficient of a question (v) measured by a specific method for a CI as the correlation between the CI of interest and the true score for the question formulated by this specific method, i.e. the validity ($v^2$) = 1 – method effect squared.
4. The overall quality of a question ($q^2$) is equal to the product of the reliability and the validity i.e. $q^2 = r^2 \times v^2$

**Designs and models for the estimation of the quality indicators**

1. Without repeated observations the reliability of survey questions cannot be determined (Lord and Novick 1969)
2. Without observation of different CI using the same method the validity and method effects or common method variances can not be determined (Campbell and Fiske (1959)
3. It follows that the only design that satisfies these criteria to determine the quality indicators of survey questions is the Multitrait Multimethod design (Andrews 1984)
4. For identification the minimal requirement is that one uses 3 CI and 2 methods. However in order to avoid empirical identification problems one better uses minimally 3 CI and 3 methods (Andrews 1984)
5. Given this design the estimation of the quality indicators can be done by the True Score MTMM model developed by Saris and Andrews (1991) which is equivalent to the classical MTMM model which provides mixed parameters (Andrews 1984)
6. In order to prevent memory effects the distance between the observation of the repeated observations have to be at least 25 minutes or 75 questions (Van Meurs and Saris (1991, p. 145)
7. In order to prevent memory effects in the third repetition it has been suggested to use a split ballot MTMM design (SB MTMM design). In this design the sample is split in more groups at random and each group is confronted with only two repetition of the questions about the same CI with a distance in-between the observations of more than 25 minutes (Saris, Satorra and Coenders 2004).
8. One will never be able to study the quality of all questions necessary in a study because one needs at least one repetition of each question to estimate the quality indicators. Although the European Social Survey (ESS) collected information about the quality of 3,000 questions over the years, there are more than 60,000 questions asked for which the quality is not determined.
9. If one knows the relationships between question characteristics and the quality of questions one can use this relationship to make predictions of the quality of any question. This means that correction for measurement errors is possible for the relationships between all variables without repeated observations of the questions. This is the fundamental idea behind the development of the program SQP (Saris et al. 2011)

**Fundamental assumptions for model specification**

1. Important with respect to the approach to estimate the quality of questions is what one thinks of the characteristic of the CI: continuous or discrete?

2. There are CI that are by definition categorical such as background variables like gender, occupation, marital status, or education.
3. Attitudinal CI are most of the time continuous because they represents judgments from extremely negative to extremely positive where all intermediate position are in principle possible
4. Because the ESS is directed to evaluations in attitudinal variables through time we will concentrate on the measurement of CI that are assumed to be continuous.

**Possible approaches to estimate the quality indicators of survey questions**

1. If the observed variables (y) are continuous or there are so many categories that is does not matter whether we treat them as continuous then the quality of the questions can directly be estimated using MTMM experiments on the basis of the Pearson correlations. We refer to this as the "one step approach".
2. For CI that are by definition continuous but are measured using questions with a limited number of categories, one can estimate the nonlinear relationship between the observed variables (y) and underlying continuous variables (y*) using nonlinear Latent trait models (Mplus). One can also correct the correlations for the categorical character of the observed variables (y) using polychric/poliserial corr (LISREL). After that the quality of these underlying continuous variables (y*) can be evaluated using for example MTMM experiments. We refer to this approach as the "two step approach".
3. If one can specify exogenous variables without measurement errors that have effect on the latent variables (y*) one could estimate the relationships between the y* variables and between the CI without any further assumptions (see Mpus). However, because we cannot imagine that there are any variables without measurement error, we will not continue with this approach.
4. For CI that are by definition categorical the most obvious solution is the use of Latent class analysis (Oberski, Hagenaars, and Saris (forthcoming)). This seems an unjustified approach for CI that are in principle continuous. Besides that it leads to very different models for the CI. Therefore we will ignore this possibility

**The assumptions of the one and two step approach and bias**

1. Applying the "one step approach" for estimation of the quality of questions which are categorical but are measures for continuous CI seems in principle incorrect but one may study how biased and inconsistent this approach is.
2. Applying the "two step procedure" assumes that the latent variables (y*) have a multivariate normal distribution. If that is not the case the approach may be biased and inconsistent. One should study these effects.

**Empirical research of the biases**

1. The possible bias of the different approaches can be studied by generating data for a set of latent (CI) and supposed continuous variables (y*) behind the observed variables (y) given a specific model, distributions of the variables and certain values of the parameters. For these data all characteristics are known. An example is given in Figure 1.

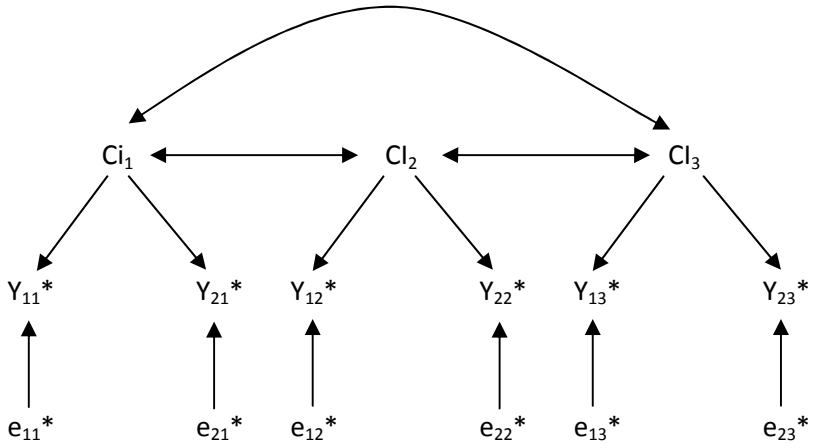


Figure 1. An example of a model used to generate data

The next step is to categorize the continuous variables (y*) in order to generate the categorical observed variables (y). This can be done in different ways with respect to: the number of categories, the size of the categories, the distribution of the categorical variables and the value attached to the different categories. This process is presented in Figure 2.

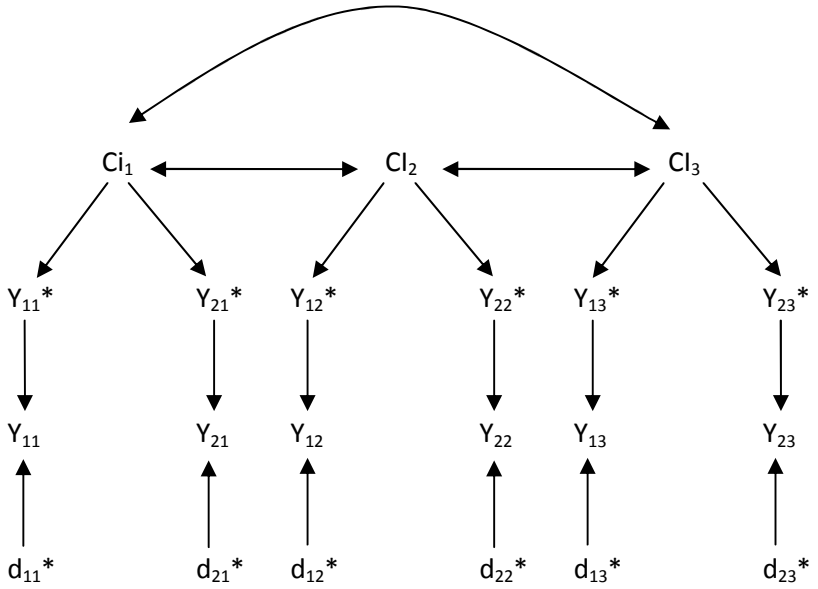


Figure 2. An example of a model used to generate data and categorize the latent continuous variables (y*) in categorical observed variables (y). In this model the disturbance ($d_{ij}$) combines the errors ($e_{ij}$) of Figure 1 and the categorization and transformation effects if the categorical characteristic of the scales is ignored.

2. This process can be repeated hundreds of times. For these data sets the characteristics and the parameter values in the population are all known. These hundreds of data sets can be analysed using the different approaches on the basis of the supposed observed categorical variables ($y_{ij}$) and in this way one gets estimates of the parameters of the model. Comparing these results with the population values one can see how much bias the different estimation procedures generate under different conditions for the different sets of parameters.
3. One can make comparison with respect to:

a. the correlations between the supposed continuous response variables (y*),
b. the quality of the different measures ($\lambda_{ij}$), and
c. the correlations between the CI ($\phi_{ij}$).

**The results of the Monte Carlo studies**

1. Results with respect to the correlations between the supposed continuous response variables (y*):
   Polychoric correlations are unbiased estimates of the correlation among the continuous variables (y*) if the variables are normally distributed (Jöreskog and Sörbom, 1988) . If the underlying variables are skewed, they can be biased (O´Brien & Homer,1987; Quiroga, 1992).However, the bias of the Pearson correlations being estimates of the correlations of the observed variable (y) and not y* is usually larger than that of the Polychoric correlations. The exception mentioned by Coenders et al. (1997) is the situation when one of the observed variables is continuous and not normal distributed. In that case the estimated correlations can be very biased.

2. Results with respect to the correlations between the CI:
   Several studies (Homer and Brien, 1988; Johnson and Creech 1983, Coender and Saris 1995) show that the one step and two step procedure perform equally well with respect to estimation of the correlations between the CI and consequentially also with respect to structural equation models specified between these variables.
   This is a very important but also surprising result if one takes into account that the correlations used in the analysis are very different. The reason can be found in the quality estimates.

   In case the variables y* are not normally distributed with low transformation errors, the one step procedure performs even better than the two steps procedure (Coenders et al. ,1997)
   On the other hand, Johnson and Crecch (1982), Olsson 1979, Quiroga (1992) and Coenders et al. (1997) suggested that sometimes correlated errors are necessary in case of the use of the one step procedure, especially when the transformation errors are very large, i.e. if the categorization is very skewed and the numbers given to the categories are equal distance (such as 1,2,3,4,5 etc).

3. Results with respect to the quality estimates:
   While the correlations on which the analyses are based in the one step and two step procedure are very different, the correlations estimated between the CI are in general very similar. This is only possible:
   1. If the estimates of the qualities of the questions are also very different, or
   2. if both approaches are equally good to correct for measurement errors.

   In the two step procedure the first step represent the correction of the correlations between the observed variables for the categorical character of these variables. The second step is then the estimation of the quality coefficients ($\lambda_{ij}$*) model of Figure 1.

In the one step procedure the model in Figure 2 is estimated but ignoring the in between level of the continuous variables. The categorization (grouping and transformation) of the variables y* will certainly not allow for a perfect relationship between the y* and the observed variables y. Therefore we can expect the quality estimates (being an indirect effect in Figure 2) will be lower than the quality estimate in Figure 1. This means that in the one step procedure the quality estimates, the effect of $CI_i$ on $y_i$ not only represent the product of the reliability and the validity but also the effect of the categorization (Oberski 2011).

Coenders et al. (1997) have shown that the size of these quality estimates in the one step procedure is in general very close to the correlations between the latent CI and the observed variables. The bias is in general smaller than .04 under many different conditions. It is for this reason that the one step procedure also for categorical variables can be used for correction for measurement errors in the relationships between CI and between CI and CP.

Both procedures have their limits. The two step procedure will not work well if the assumption of multivariate normal distribution does not hold, an assumption that is difficult to test (Quiroga 1992). The one step procedure will not work well if the transformation errors are very large which can make the relationships between the observed variables very nonlinear.

**The effect of the estimation procedure in MTMM experiments and meta analyses**

The above quoted literature used for the tests of the procedures only simple multiple indicator models. We are normally using MTMM experiments which make a distinction between the reliability and validity (Saris and Andrews, 1991). So the question is also what is the effect on the reliability and validity estimates of the categorization? Besides that one can ask the question if these effects change the results of meta analysis across sets of MTMM experiments?

In an international research project (Saris et al. 1996) with respect to the evaluation of the MTMM approach two studies have been done using MTMM experiments. One is a Monte Carlo studies like the ones that we have discussed above (Költringer 1995) and the other is a study using real data (Saris, Van Wijk en Scherpenzeel, 1998). The Monte Carlo study of Költringer confirmed the results for the MTMM experiments that were found by the previous Monte Carlo studies. New was that this study showed that only the reliability coefficients of the True score MTMM model were affected by the categorization and not the validity coefficients.

The study using real data was based on MTMM experiments in 10 different countries where in total 17 studies were done. In all countries the same questions were used with respect satisfactory with life in general, housing, financial situation and social contacts. In all countries a 100 point, 10 point and a 5 point scale were used but in some studies other scales were added like a 4 point scale and Grafical line scale or, as called nowadays, Visual Analog or VAS scale. The basic description of the designs of the different experiments is presented in Table 1. The table shows that the order of the scales varied by country and some experiments were

done in a cross sectional study others were done in a panel with two methods at each occasion. Table 1 also shows that the data collection method from study to study. For more details about all these experiments we refer to Saris et al. (1996).

**Table 1. Design differences between the countries.**

| Country and study | Order of the scales | | | | Number of interviews | Data-collection technique | N |
|---|---|---|---|---|---|---|---|
| Slovenia | 100 | 10 | 5 | | 1 | face-to-face | 2050 |
| Germany | 100 | 10 | 10 | 5 | >1 | telephone | 209 |
| Spain: Catatonia | 100 | 5 | 10 | | 1 | telephone | 406 |
| Italy | 5 | 10 | 100 | | 1 | face-to-face | 1010 |
| Belgium: Flanders | 5 | 10 | 100 | | 1 | personal | 624 |
| Belgium: Wallone | 5 | 10 | 100 | | 1 | personal | 439 |
| Belgium: Brussels | 5 | 10 | 100 | | 1 | personal | 376 |
| Norway | 10 | 5 | line | 10 | >1 | telephone + mail | 231 |
| Sweden: Göteborg area[a] | | varied | | | >1 | mail | 336 |
| Netherlands, study 1 | 10 | 100 | 5 | | >1 | telepanel | 486 |
| Netherlands, study 2a | line | 10 | 100 | 5 | >1 | telepanel | 1599 |
| Netherlands, study 2b | 5 | 100 | | | 1 | mail | 1219 |
| Netherlands, study 3 | 5 | 10 | 4 | 100 | >1 | telephone | 424 |
| Russia: Russians | 100 | 10 | 5 | | 1 | face-to-face | 7671 |
| Russia: Tartars | 100 | 10 | 5 | | 1 | face-to-face | 848 |
| Russia: Other nationalities | 100 | 10 | 5 | | 1 | face-to-face | 1502 |
| Hungary | 5 | 10 | 100 | | 1 | mail | 300 |

For all experiments the data have been analyzed with the one step (pearson correlations) and the two step procedure (polychoric and polyserial correlations). Before presenting the results for all countries it is interesting to look at a typical example. For this purpose we have chosen study 1 in the Netherlands where the data were collected using an early version of a Web survey (telepanel) using a panel design. The sample consisted in this case of 486 persons randomly chosen from the Dutch population.

In Table 2 and 3 the correlations between the observed variables are presented. In Table 2 the correlations have been calculated using polychoric and polyserial correlation coefficients. In Table 3 the standard Pearson correlations are computed. If we look at the matrices in bold, which represent the correlations between the different variables for each method, we see no clear effect of the number of categories in Table 2. This can be expected because the correlations are corrected for categorization effects.

**Table 2. The Polychoric / Polyserial correlations between the 12 satisfaction measures**

| | | 100 p | | | | 10 p | | | | 5 cat | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | sat life | sat hous | sat fina | sat cont | sat life | sat hous | sat fina | sat cont | sat life | sat hous | sat fina | sat cont |
| 100 p | sat life | 1.000 | | | | | | | | | | | |
| | sat hous | **.382** | 1.000 | | | | | | | | | | |
| | sat fina | **.508** | **.403** | 1.000 | | | | | | | | | |
| | sat cont | **.468** | **.281** | **.281** | 1.000 | | | | | | | | |
| 10 p | sat life | .493 | .342 | .354 | .353 | 1.000 | | | | | | | |
| | sat hous | .306 | .716 | .251 | .225 | **.510** | 1.000 | | | | | | |
| | sat fina | .399 | .367 | .722 | .253 | **.544** | **.398** | 1.000 | | | | | |
| | sat cont | .381 | .282 | .233 | .647 | **.569** | **.442** | **.413** | 1.000 | | | | |
| 5 cat | sat life | .525 | .335 | .369 | .387 | .535 | .322 | .403 | .410 | 1.000 | | | |
| | sat hous | .352 | .675 | .274 | .209 | .329 | .695 | .286 | .251 | **.476** | 1.000 | | |
| | sat fina | .432 | .319 | .692 | .243 | .309 | .242 | .732 | .203 | **.510** | **.414** | 1.000 | |
| | sat cont | .386 | .203 | .116 | .570 | .315 | .218 | .201 | .646 | **.524** | **.319** | **.289** | 1.000 |

**Table 3. The Pearson correlations between the 12 satisfaction measures**

| | | 100 p | | | | 10 p | | | | 5 cat | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | sat life | sat hous | sat fina | sat cont | sat life | sat hous | sat fina | sat cont | sat life | sat hous | sat fina | sat cont |
| 100 p | sat life | 1.000 | | | | | | | | | | | |
| | sat hous | **.382** | 1.000 | | | | | | | | | | |
| | sat fina | **.508** | **.403** | 1.000 | | | | | | | | | |
| | sat cont | **.468** | **.281** | **.281** | 1.000 | | | | | | | | |
| 10 p | sat life | .467 | .336 | .337 | .337 | 1.000 | | | | | | | |
| | sat hous | .287 | .702 | .236 | .209 | **.456** | 1.000 | | | | | | |
| | sat fina | .390 | .361 | .708 | .245 | **.505** | **.355** | 1.000 | | | | | |
| | sat cont | .366 | .282 | .222 | .629 | **.524** | **.398** | **.375** | 1.000 | | | | |
| 5 cat | sat life | .463 | .304 | .323 | .341 | .447 | .262 | .340 | .344 | 1.000 | | | |
| | sat hous | .325 | .641 | .251 | .191 | .287 | .606 | .250 | .219 | **.381** | 1.000 | | |
| | sat fina | .411 | .303 | .650 | .231 | .272 | .208 | .659 | .178 | **.428** | **.351** | 1.000 | |
| | sat cont | .356 | .186 | .105 | .527 | .271 | .183 | .174 | .556 | **.433** | **.262** | **.246** | 1.000 |

In Table 3 we see that the 5 point scale produces clearly lower correlations than the other two methods while it seems that the 10 point scale generates even higher correlations than the 100 point scale. This can be due to unequal method effects I these scales or larger random errors in the 100 point scale. This cannot be determined without looking at the estimated quality indicators for these measures.

Table 4 presents the estimates of the quality indicators for all measures estimated by the two step (part a) and one step procedure (part b).

**Table 4. Quality Estimates for the 12 Satisfaction Measures**

| | Reliability coefficients | | | | Validity coefficients | | | | Method effects | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | life | hous | fina | cont | life | hous | fina | cont | life | hous | fina | cont |
| a. For polychoric/polyserial correlations | | | | | | | | | | | | |
| 100-p scale | 0.786 | 0.908 | 0.876 | 0.828 | 0.930 | 0.948 | 0.944 | 0.937 | 0.367 | 0.318 | 0.330 | 0.349 |
| 10-p scale | 0.803 | 0.914 | 0.970 | 0.949 | 0.856 | 0.891 | 0.904 | 0.899 | 0.517 | 0.454 | 0.428 | 0.437 |
| 5-cat scale | 0.821 | 0.899 | 0.907 | 0.818 | 0.900 | 0.918 | 0.919 | 0.900 | 0.435 | 0.397 | 0.394 | 0.437 |
| b. For Pearson correlations | | | | | | | | | | | | |
| 100-p scale | 0.785 | 0.936 | 0.878 | 0.833 | 0.938 | 0.957 | 0.951 | 0.946 | 0.345 | 0.290 | 0.308 | 0.325 |
| 10-p scale | 0.762 | 0.865 | 0.935 | 0.906 | 0.854 | 0.889 | 0.906 | 0.899 | 0.520 | 0.458 | 0.424 | 0.437 |
| 5-cat scale | 0.726 | 0.825 | 0.846 | 0.745 | 0.889 | 0.915 | 0.920 | 0.895 | 0.458 | 0.403 | 0.393 | 0.446 |

Table 4 shows that the quality Indicators for the 100 point scale are very similar for the two methods. For the 10 point scale the reliability coefficients are already considerably lower for the Pearson correlations that for the Polychoric correlations while the validity and method effects are comparable. For 5 point scale the reliabilities for the Pearson correlations are even lower and much lower than the Polychoric correlations. On the basis of the Monte Carlo experiments these results were to be expected because the polychoric correlation coefficient corrects for categorization and the Pearson correlation does not so in that case the "reliability coefficients" include also the categorization effects.

The literature also predicts that we should get approximately the same estimated correlations between the variables corrected for measurement errors. How this correction is done has been discussed in Bollen (1989) using latent variable models and Saris and Gallhofer( 2007/ 2014) and Saris and DeCastellarnau (forthcoming) using a simpler approach. Table 5 shows indeed that the correlations between the CI are very similar. Give that the correlation in table 2 and 3 were very different, the quality estimates are so different that this difference, due to the categorization effect, compensates for these differences.

**Table 5. Estimated Correlation Matrix of the Trait Factors**

| Polychoric/polyserial | | | | | Pearson | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | LIFE | HOUS | FINA | CONT | | LIFE | HOUS | FINA | CONT |
| LIFE | 1.000 | | | | LIFE | 1.000 | | | |
| HOUS | 0.550 | 1.000 | | | HOUS | 0.542 | 1.000 | | |
| FINA | 0.630 | 0.406 | 1.000 | | FINA | 0.637 | 0.402 | 1.000 | |
| CONT | 0.667 | 0.347 | 0.309 | 1.000 | CONT | 0.670 | 0.342 | 0.310 | 1.000 |

Finally we like to represent the meta-analysis across all MTMM experiments taking into account all the characteristics in which the experiments varied. Table 6 presents the results of the analysis for the reliability and the validity coefficients obtained from the Polychoric/Polyserial correlations (PPC) and the Pearson correlation coefficient (PCC).

**Table 6. The estimated effects of instrument characteristics on quality estimates based on the Polychoric/Polyserial (PPC) and Pearson correlation coefficients(PCC)**

| | | Validity Coefficient | | Reliability Coefficient | |
|---|---|---|---|---|---|
| | | PPC | PCC | PPC | PCC |
| | | Mean=.940 | Mean = .94 | Mean=.911 | Mean = .879 |
| measures | N | Deviations | Deviations | Deviations | Deviations |
| SATISFACTION DOMAIN | | | | | |
| Life in general | 54 | -.006 | -.006 | -.038 | -.043 |
| House | 54 | .005 | .004 | .029 | .033 |
| Finances | 54 | .003 | .003 | .020 | .025 |
| Social contacts | 54 | -.001 | -.002 | -.011 | -.015 |
| RESPONSE SCALE | | | | | |
| 100 p. number scale | 64 | -.021 | -.018 | -.027 | .015 |
| 10 p. number scale | 72 | .011 | .011 | .051 | 049 |
| 5/4 p. category scale | 72 | .001 | -.001 | -.026 | -.067 |
| graphical line scale | 8 | .058 | .058 | -.007 | .038 |
| DATA COLLECTION | | | | | |
| Face-to-face interview | 96 | .011 | .010 | .012 | .004 |
| Telephone interview | 52 | .002 | -.001 | -.051 | -.046 |
| Mail questionnaire | 40 | -.014 | -.014 | -.011 | -.003 |
| Telepanel interview | 28 | -.022 | -.015 | .067 | .075 |
| POSITION | | | | | |
| 1 - 5 | 48 | .011 | .017 | .026 | .019 |
| 6 - 49 | 68 | .017 | .017 | -.001 | -.005 |
| 50+ | 100 | -.017 | -.020 | -.012 | -.006 |
| TIME BETWEEN REPETITIONS | | | | | |
| alone in interview | 32 | .010 | .006 | -.071 | -.070 |
| first/last 5-20 minutes | 64 | .017 | .022 | .063 | .067 |
| first/last 21-60 minutes | 80 | -.021 | -.025 | -.023 | -.027 |
| middle, 5-20 minutes | 16 | .043 | .049 | .028 | **.051** |
| middle, 21-60 minutes | 24 | -.017 | -.017 | -.016 | **-.030** |
| ORDER OF PRESENTATION | | | | | |
| first measurement | 60 | -.015 | -.016 | -.025 | -.025 |
| repetition | 156 | .006 | .006 | .010 | .010 |
| COUNTRY | | | | | |
| Slovenia | 12 | .020 | .025 | -.013 | .009 |
| Germany | 16 | .007 | .019 | .028 | .025 |
| Catalonia (Spain) | 12 | -.039 | -.045 | -.022 | -.022 |
| Italy | 12 | .013 | .024 | .043 | .056 |
| Flanders (Belg)+ Netherlands | 64 | -.028 | -.034 | -.039 | -.049 |
| Wallonia (Belgium) | 12 | -.026 | -.031 | -.028 | -.033 |
| Brussels (Belgium) | 12 | .006 | .003 | .000 | -.012 |
| Sweden | 12 | .023 | .029 | .099 | .090 |
| Hungary | 12 | .050 | .050 | .046 | .054 |
| Norway | 16 | -.018 | -.018 | .031 | .023 |
| Russians (Russia) | 12 | .043 | .046 | .004 | .022 |
| Tartarians (Russia) | 12 | .033 | .037 | .003 | .018 |
| Other nationalities in Russia | 12 | .039 | .042 | .000 | .014 |
| Multiple $R^2$ | | .331 | .345 | .616 | .688 |

The striking result is that all effects of the different characteristics of the experiments of the quality coefficients are approximately the same except the effect of the number of categories of the scales. We have seen that for the one data set we have illustrated before but here we see this confirmed for all the other data sets as well.

This result shows that the predictions of the quality of questions on the basis of such meta analysis like above and also in SQP 2.0 will be different if they are based on the correlations based on the polychoric correlations coefficients or the Pearson correlation coefficients. The two steps procedure gives the quality estimates (reliability and validity) for the continuous variables behind the observed variables after correction for the categorization errors. The one step procedure provides the same quality estimates for the questions asked but including the categorization effect.

**Conclusions**

1. Correction for measurement errors is necessary in order to study the distribution and relationships between the CI, and between CI and CP and to compare relationships across countries.
2. In general the two and one step approach both are able to correct for measurement errors and generate very similar relationships between the CI which means that all further analyses based in this information will be the same as well
3. The two step procedure provides estimates of the quality of the continuous variables (y*) behind the observed variables.
4. The one step procedure provides estimates of the quality of the observed variables (y).
5. These estimates are so different that the estimates based on the two step procedures can only be used if the polychoric /polyserial correlations are used to estimate the correlations between the variables. The estimates based on the one step procedure can only be used if Pearson correlations are used to estimate the correlations between the observed variables.
6. The two step procedure provide the estimates of the reliability and validity of the continuous variable (y*) behind the observed variables. The one step procedure provides an estimate of the same validity but the so called “reliability” is based on a combination of random errors and categorization errors in the observed variable (y).

Are there reasons to prefer the one approach above the other?

1. If one want to obtain the quality of the observed variables (y) , including the categorization effects, one should use the one step procedure.
2. If one wants to obtain the quality of the continuous variables (y*) one should use the two step procedure.
3. The one step procedure provides biased results if there are serious transformation errors in the observed variables. This happens if the categorization of the observed variables is very skewed and rank numbers are provided to the categories. This problem can be prevented by making more regular sized categories.

4. The two step procedure provides biased results if the latent variables behind the observed variables do not have a multi-normal distribution. The biases are especially serious in case of non-normally distributed continuous variables between the observed variables.
5. In the social sciences often composite scores are used for the CP based on weighted or unweighted averages of the observed variables. This approach allows for very simple models for further analysis (Saris and Gallhofer 2014). These composite sores contain also measurement errors. The quality should be known to correct for measurement errors. On the basis of the quality of the observed variables (y) the quality of the composite scores can easily be computed as shown by Saris and Gallhofer (20014) using the quality of the questions estimated with the one step procedure. How this can be done using the two step procedure is not at all clear.
6. The two step procedure also allows further analysis between the CP or CP and CI but then complex latent variables models have to be used.
7. Both approaches are equally good for equivalence testing after correction for measurement errors. In this context we refer to Saris and Gallhofer (2014).

**References**


Andrews, F.M. (1984). Construct validity and error components of survey measures: a structural modelling approach. *Public Opinion Quarterly*, 48(2), 409-422.

Andrews, F.M., Morgan, J.N., Sonquist, J.A. and Klem, L. (1973). *Multiple classification analysis*. Ann Arbor, MI, Institute for Social Research.

Bollen, K.A. (1989). *Structural equations with latent variables.* New York: Wiley.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multimethod-multitrait matrix. *Psychological Bulletin*, 56(2), 81-105.

Coenders G. and W. E. Saris (1995). Categorization and measurement quality. The choice between Pearson and Polychoric correlations. In Saris and Münnich (Eds.). *The Multitrait-Multimethod approach to evaluate measurement instruments*. Budapest, Eötvös University Press, Chapter 7.

Coenders G. , A. Satorra and W.E.Saris (1997 ). *Alternative Approaches to Structural Modeling of ordinal data: A Monte Carlo Study*. Structural Equation Modeling, 4(4), 261-282.

Homer,P and R.M.O'Brien (1988). Using Lisrel models with crude rank category measures, Quality and Quantity, 22(2), 191-201.

Hoogendoorn, A. (1994). Note on and program for the derivation of standard errors for MCA coefficients. NIMMO, University of Amsterdam.

Johnson D.R. and Creech J.C. (1983). Ordinal measures in multiple indicator models: a simulation study of categorization error. American Sociological Review, 48(3), 398-407.

Jöreskog, K.G. and Sörbom, D. (1988). *Prelis: a program for multivariate data screening and data summarization* (second edition). Mooresville, Scientific Software.

Költringer, R. (1995). Categorization and measurement quality: a population study using artificial multitrait-multimethod data. In Saris and Münnich (Eds.). *The Multitrait- Multimethod approach to evaluate measurement instruments*. Budapest, Eötvös University Press, Chapter 6.

Költringer, R. (1993). Messqualität in der sozialwissenschaftlichen Umfrageforschung. Endbericht Project P8690-SOZ des Fonds zur Förderung der wissenschaftlichen Forschung (FWF), Wien.

Lodge M. (1981). Magnitude scaling: quantitative measurement of opinions. Sage University press Series: Quantitative Applications in the Social Sciences. Beverly Hills: Sage.

Lord, F. and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA, Addison-Wesley.

Northrop F.S.C. (1947). The logic of the sciences and the Humanities, New York: The Macmillan Company.

Oberski D. (2011) Latent Class Multitrait-Multimethod Models. In Oberski, D. *Measurement error in comparative research*. Doctoral Dissertation, University of Tilburg.

D.L. Oberski, J. Hagenaars, W.E. Saris, (forthcoming), "The Latent Class Multitrait-Multimethod Model", *Psychological Methods*.

O´Brien R.M. and P.Homer (1987). Corrections for coarsely categorized measures: LISREL´s polyserial and Polychoric correlations. *Quality and Quantity*, 21(4), 349-360.

Ollsen U. (1979). Maximum Likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443-460.

Quiroga A.M. (1992) *Studies of the Polychoric Correlation and Other Correlation Measures for Ordinal Variables*. Doctoral dissertation, University of Uppsala.

Saris, W.E. (1990). The choice of a research design for MTMM studies. Chapter 8 in: Saris, W.E. and van Meurs, A. (Eds.). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. Amsterdam, North Holland.

Saris, W.E. and Andrews, F.M. (1991). Evaluation of measurement instruments using a structural modelling approach. In Biemer, P.P., Groves, R.M., Lyberg, L.E. Mathiowetz, N. and Sudman, S. (Eds.). *Measurement errors in surveys*. New York, Wiley and Sons.

Saris, W.E. and van Meurs, A. (Eds.) (1990). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. Amsterdam, North Holland.

Saris W.E. and A.Münnich (Eds.) (1995) *The Multitrait-Multimethod approach to evaluate measurement instruments*. Budapest, Eötvös University Press.

Saris, W.E., Veenhoven, R., Scherpenzeel, A.C. and B.Bunting (Eds.) (1996) *Life satisfaction in Western and Eastern Europe*.Budapest. Eötvös University Press.

Saris W.E., T.van Wijk and A. Scherpenzeel (1998) Validity and reliability of subjective social indicators: the effect of different measures of association. Social Indicators Research, 45(1-3), 173-199

Saris W. E., A. Satorra, and G. Coenders 2004. A New Approach for Evaluating the Quality of Measurement Instruments: The split-Ballot MTMM Design. *Sociological Methodology*, 34(1), 311─347.

Saris W.E. and I.N. Gallhofer (2007) Design, evaluation and analysis of questionnaires for survey Research, Wiley , Hoboken.

Saris W.E. and I.N. Gallhofer (2014) Design, evaluation and analysis of questionnaires for survey Research (second edition) , Wiley, Hoboken.

Saris W. , D. Oberski, M. Revilla, D. Zavala, L. Lilleoja, I.Gallhofer and T. Gruner (2011) The development of the program SQP 2.0 for the prediction of the quality of survey questions, RECSM Working paper 24, Universitat Pompeu Fabra, Barcelona.

Saris W.E and A. de Castellanau (forthcoming 2014) Correction for measurement errors. *ESS/EduNet*

Van Meurs, A. and Saris, W.E. (1990). Memory effects in MTMM studies. Chapter 6 in: Saris, W.E. and van Meurs, A. (Eds.). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. Amsterdam, North Holland.

Van Wijk T. (1996). Quality and Correlation: An evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies. Amsterdam, Unpublished master thesis of the University of Amsterdam.